Re-orderable Matrix and Parallel Coordinates Applied to Visualizing Data on Substitute Teaching

Janette Cheng
Stanford University
janette.cheng@stanford.edu

Gloria Chua
Stanford University
cjmg@stanford.edu

ABSTRACT

In this paper, we describe the design and implementation of a data visualization application around data related to the state of substitute teaching in the United States. This web-based visualization effectively serves two user groups – prospective substitute teachers and the general public – by supporting both targeted state-level navigation, as well as more exploratory interactions. The primary features of the web-based visualization are re-orderable matrices, parallel coordinates and state information bars for details. While previous work has explored combining re-orderable matrices with parallel coordinates for a single data set, we do so with two different data sets. Preliminary user testing demonstrates the effectiveness of this visualization in helping users understand the complexity within this niche domain of substitute teaching.

INTRODUCTION

Substitute Teaching Data

Among the many topics within the realm of education reform, the topic of substitute teaching is often neglected. There is limited press around the topic, and public information is limited to text-heavy governmental documents that are difficult to understand at first glance. [1, 2] This presents a novel application for data visualization for a particularly niche dsata set.

Despite this, 10% of a student's time in the classroom is spent with a substitute teacher [3]. The presence of substitute teachers is crucial in maintaining and improving the quality of regular teachers as well – teachers are entitled to take time off for personal reasons, sick leave, or professional development, and substitute teachers are necessary to keep the classroom curriculum going. $4 billion (1% of the US Education Budget) is spent annually on substitute teaching [4], but little is publicly understood about the intricacies of substitute teaching, including teacher salary, requirements necessary to become a substitute teacher, and how these factors differ from state to state. Neither is substitute teaching understood in the context of the broader education system through state-by-state comparisons.

Figure 1. Existing data on substitute teaching: Screenshots of selected Bay Area school district websites' substitute teaching pages (San Francisco, Sunnyvale, Hayward, Oakland, Palo Alto, Emery)

Target Audience

There are two main target audiences who might find understanding the intricacies of substitute teaching helpful. The first group is prospective substitute teachers, and the second group is a general audience interested in learning about topics in public education.

Our visualization aims to fulfill the needs of both audiences. In the Methodology section, we will describe the design principles and features that were employed to successfully achieve user goals.

RELATED WORK

Preliminary Design Choices

The main graphical components of our visualization consist of an interactive matrix to encode substitute teaching requirements, and a set of parallel coordinates to encode quantitative data such as teacher pay and reading/math scores for each state. One of the goals of our final visualization is to be able to identify larger trends within substitute teaching requirements, specifically similarities between states, and to tie these trends to quantitative measures. Thus, the majority of the past work we looked at was concerned either with re-orderable matrices or parallel coordinates.

Information on Substitute Teaching

Currently, two types of information on substitute teaching are available. The first, whose target audience is prospective substitute teachers, are school district websites that provide information on the steps and requirements to be a substitute teacher, and many of these school districts repeat similar state requirements, in formats that are text-heavy and difficult to navigate. (Figure 1 displays screenshots of such websites of selected school districts in the San Francisco Bay Area for reference.)

The second type is macro-level studies on substitute teaching, meant for a specialized audience, like a National Education Association study [1] or text-heavy research papers [2] that are difficult to scan.

No known data visualization application for a general layperson audience has been created around this topic at the time of publication.

Re-orderable Matrices

Our work builds upon Jacque Bertin's re-orderable matrices - data tables where numeric cell values have been replaced with appropriately sized ink blobs. [5] The most important aspect of this particular visualization is that the viewer is capable of rearranging rows in order to group together rows that are visually similar, a task Bertin claims is natural and doesn't require special skills.

Re-orderable matrices have been shown to be particularly suitable for pattern discovery in multi-dimensional applications, such as in archaeology [6] and ecology [7].

We build on prior work that adds the ability to automatically group similar rows (manual grouping is time-consuming, especially for novice users). Siirtola and Mäkinen [8] suggest a "barycentric heuristic," which given a matrix with binary cell values, treats it as an adjacency matrix and reorders the vertices in the resulting bipartite graph based on the averages of their adjacent vertices in the opposite vertex set. In two user studies, one with 12 and the other with 16 participants, it was found that this "correlation coefficient" view resulted in an approximate 60% improvement in time and accuracy over the classic manual user interface for the 16-participant study, and around 24% improvement in the 12-participant study.

Substitute teaching requirements are multivariate and each requirement is binary, therefore they are suitable to visualize using re-orderable matrices.

Parallel Coordinates

Our visualization also builds on prior work in parallel coordinates. We borrow general principles established by Schneiderman on task taxonomy for high-level interaction design (Overview, Zoom, Filter, Details-on-demand, History, Extract) [9] in designing our parallel coordinates.

Siirtola and Raiha point out key issues to consider while designing our parallel coordinate interactions [10]. They include obscured data relationships due to non-adjacencies of coordinate axes, and occlusion of data itself due to overlapping lines. Siirtola and Raiha also suggest a conventional technique of multi-dimensional brushing to select and filter data, and a hierarchical clustering method to abstract closely related data lines into a thicker one to address the occlusion problem. We kept these issues in mind while implementing our system.

Combining Re-orderable Matrices with Parallel Coordinates

Further work has also been done in linking these two data visualization structures in a single scheme. While re-orderable matrices are good for identifying groups and classes within the data, correlations and anomalies are better displayed through parallel coordinates. Another user study by Siirtola [11] demonstrate that participants also generally found the re-orderable matrix easier to use, and favored using the re-orderable matrix as the input and confirming their observations by checking the parallel coordinates view.

Although we are looking to display different data sets in our visualization (parallel coordinates for quantitative data and a re-orderable matrix for qualitative requirement data), the views can still be linked in specific ways that would enable this sort of workflow.

METHODOLOGY

Designing User Goals

As described in the introduction, our visualization has two main groups of audiences. We began by analyzing user goals of each group.

The first group is prospective substitute teachers, who might be most interested in a targeted understanding of the requirements and context for their specific state, and figuring out next steps (e.g. how to go about applying to be a substitute teacher).

The second group is a general audience who might be interested in public education, who might come across this particular visualization perhaps embedded in a news

publication. This group is more likely to want to explore the data and pick up broader trends and relationships in the topic of substitute teaching. A notable sub-group is policymakers who might find these trends useful in informing policy decisions. Some questions they might be interested in include (but are not limited to):

- Which states have the most stringent requirements to become a substitute teacher, and how might stringency be related to academic performance?
- How might a state's substitute teacher salary be related to the socioeconomic context of a state, if at all?

Visualization Elements

We note here that the first user group is likely to prefer a more target visualization with state-specific details, while the second user group is likely more interested in macro-level trends.

We devised three main modes of data visualization that enables these contrasting user needs to be fulfilled.

- *Re-orderable matrix*: displays information regarding substitute teaching requirements for different types of substitute teachers (there may be more than 1 per state)
- *Parallel coordinates*: displays state-level contextual information around substitute teaching (e.g. mean annual salary, student academic performance, poverty level)
- *Information bar*: displays state-specific details for both requirements and contextual information, and actionable steps for prospective substitute teachers

Data Collection

Data on substitute teaching requirements were provided by Audra Hollingfield on behalf of Substantial, a non-profit initiative looking at ways to improve the recruitment, training and support of substitute teachers nationwide. The data has been curated from individual state websites and compiled by Hollingfield.

Data on substitute teacher salaries came from the Bureau of Labor Statistics [12]. Reading and math scores of 4th and 8th graders are used as proxy indicators for student academic performance, and the data was taken from the National Assessment of Educational Progress [13]. Poverty level was represented by the percentage of students eligible for free and reduced lunch, provided by the National Center for Education Statistics [14]. All data is taken from the year of 2015.

Low to High Resolution Mockups

Based on the aforementioned user goals, we devised a low-resolution wireframe visualization to outline and explain our visualization elements. Figure 2 shows the low-resolution wireframe.

After displaying and explaining it to an audience of 30-40 (during class), we noted a few key points of improvement the audience raised:

- "Be sure that your visualization does not become cluttered by an overwhelming amount of data."
- "In terms of interaction points, I'd love to be able to choose locations and get data based on that because the requirements will be vastly different for each state."
- "I'd love to see you identify use cases of your data and design to support task completion."

We recognized that it was necessary to incorporate a main message in the data visualization – otherwise the visualization might not have any targeted goal or purpose, especially for the second audience (general public). Notably, we used the title element to achieve this aim: instead of a more neutral phrasing e.g. "Visualization of Substitute Teaching Data", we chose one that passed a normative evaluation on the current state of substitute teaching: "The (Ungreat) State of Substitute Teaching", pointing the reader to discover what makes the current situation not great.

Figure 3 shows the high-resolution mockups our final application is based on.

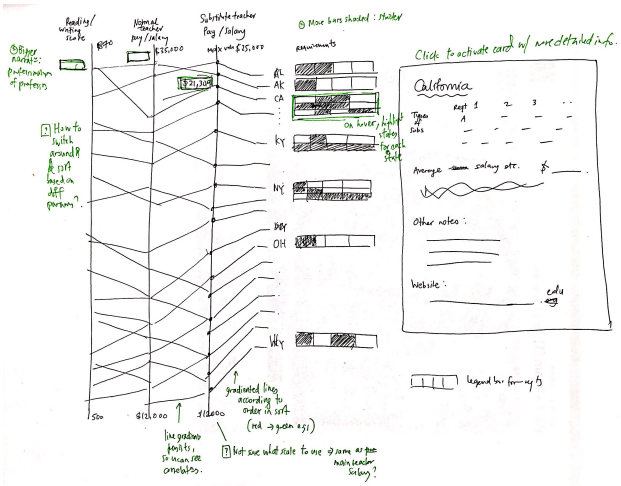


Figure 2. Low-resolution wireframe illustrating key visualization elements

SYSTEM DESCRIPTION

We chose to implement the visualization in an interactive web-based application, supported by the D3 data visualization libraries. We extracted the requirements data by hand from Word documents. The requirements and contextual data were then and converted into .csv and .json files which are accessed in the JavaScript code. For the re-orderable matrix, we looked at Mike Bostock's "Les Misérables Co-occurrence" D3 example [15] as a starting point; For the parallel coordinates, we looked at Jason Davie's D3 "Parallel Coordinates," [16] and added multiple new features to both libraries.

Figure 4 displays selected components of the final visualization. We successfully implemented all 3 visualization elements, with minor tweaks to interaction modes.

Sample user flow for prospective substitute teachers

A prospective substitute teacher is likely to be most interested in her particular state, and click directly on her state in the State column. She will be able to see a highlighted line with detailed tooltips in the parallel coordinate space, and will see highlighted bars for the various substitute teacher requirements in the re-orderable matrix.

Simultaneously, the information bar for her state appears at the bottom, displaying the specific types of permits she can apply for, as well as neatly displays the state contextual information (e.g. annual substitute teacher salary). It takes only 1 step to get to this particular end state.

Sample user flow for general public

For the general public, whose intent is likely more exploratory in nature, there are multiple ways to discover

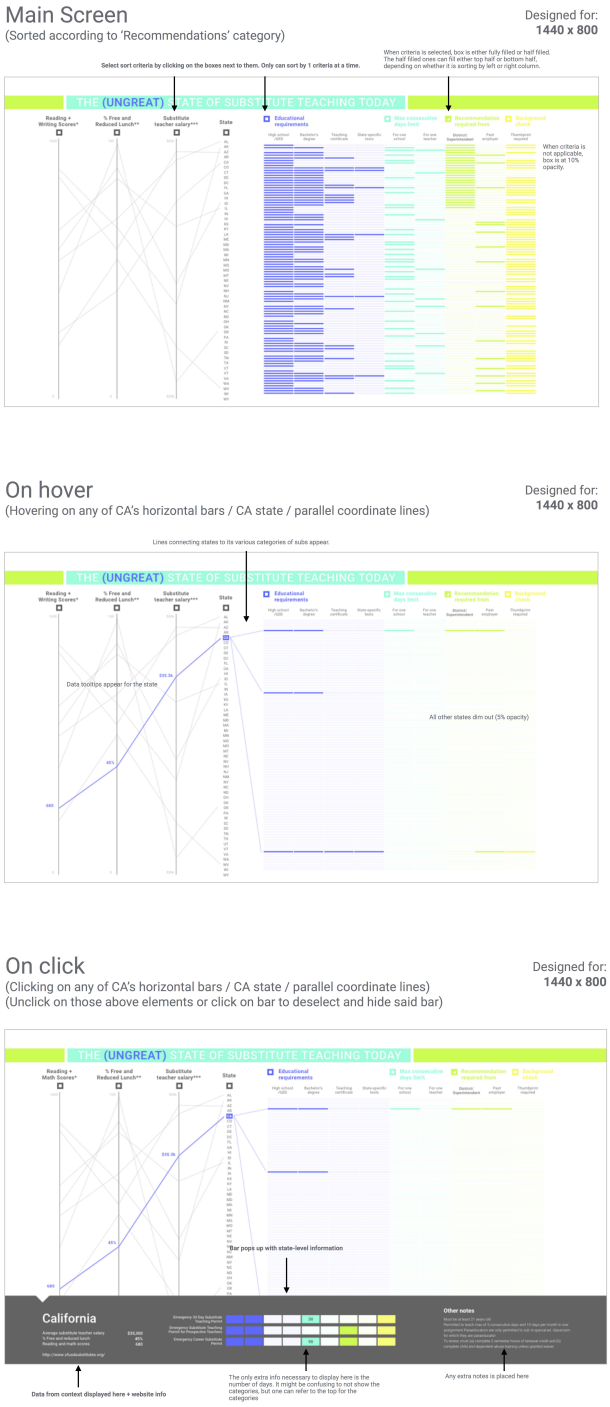


Figure 3. High-resolution mockups used to guide development

relationships in the data. In this section, we will explain the features that enable them to do so.

Re-orderable Matrix for Substitute Teaching Requirements

Each row in the matrix represents the requirements for a particular substitute-teaching permit. (Each state may have

more than one permit.) The re-orderable matrix consists of 9 different binary variables in 4 categories of requirements. First, educational requirements (4 variables: whether high school diplomas, bachelor's degrees, teaching certificates or additional state tests are necessary to become a substitute teacher). Second, whether there are a maximum number of consecutive days a substitute teacher can cover for a school, or a teacher (2 variables). Third, whether recommendations are required in a substitute teacher application from either the district, or a previous employer (2 variables). Lastly, whether there are more stringent requirements for the customary background check, be it a thumbprint requirement or an additional background check by the school (2 variables).

While there were some variables that could be more ordinal / quantitative in nature (e.g. number of consecutive days instead of whether a state has a limit on the number of consecutive days), we chose to use only binary nominal variables for consistency and ease of understanding, leaving quantities as a detail revealed only through deeper exploration.

Users are able to organize the matrix by clicking on any of the columns. For example, if a user clicks on "Bachelor's Degree" all of the rows that have a BA requirement will move up while those that do not have a BA requirement will move to the bottom of the matrix.

Users can also click on any row to highlight that row (all other rows in the matrix will turn grey) as well as the line in the parallel coordinates display that corresponds to the state. An info bar at the bottom of the page will also pop up with more specific information about the selected state's requirements (for example, if there is a limit on the maximum consecutive teaching days, the specific number of days can be found in the info bar).

Users can also use "Shift + click" to brush-select multiple contiguous rows in the matrix, which will highlight those rows and all corresponding states in the parallel coordinates display. This can be particularly useful if the user wants to see which states do/do not have a specific requirement (user can sort by requirement and then select a chunk of the rows).

Parallel Coordinates for Contextual Information

The parallel coordinates section displays 7 state-level variables around 4 categories: state name, mean annual salary (for both substitute teachers and regular teachers), student academic performance (reading and math scores for 4^{th} and 8^{th} grade), and poverty level (% population on free and reduced lunch). We selected these categories based on brief conversations with 2 education experts about what they think would be most relevant.

The parallel coordinates are ordered by category from right to left. This arrangement places the State variable in the center of the visualization, as a connecting bridge between the two sections, unifying the visualization. The specific order of state-salary-poverty-test scores is based on the contextual relevance of each category to the other. (e.g. teacher salary is more relevant to poverty than test scores). 2 variables (substitute teacher salary and % free and reduced lunch) had their axes flipped (to be in ascending instead of descending order) to visually reinforce the relationship between those variables and their adjacent ones.

In terms of interactions, we added the ability to click on specific lines in order to highlight and display number labels on each axis for a given state. The user can click on the state labels on the right-most axis or any of the lines themselves to highlight a state and all of its corresponding rows in the matrix. Clicking on either state or line will reveal the information bar below.

Asymmetric Selection between Re-orderable Matrix and Parallel Coordinate Variables

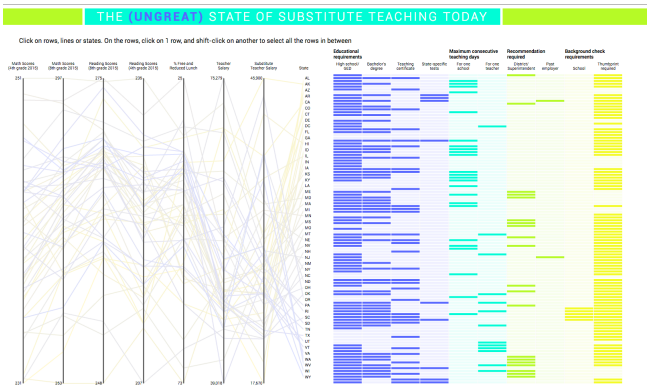
It is important to note here that selection is purposefully asymmetric between the two displays - there is a many-to-one correspondence between parallel coordinate data to re-orderable matrix data. Multiple rows in the matrix may correspond to 1 state (since each state may have more than 1 permit available), while 1 line in the parallel coordinate space directly coordinates with 1 state.

Specifically, selecting a state in the parallel coordinates display may result in several highlighted matrix rows; however, clicking on a row in the matrix will only highlight that particular row and its corresponding parallel coordinates line (will not highlight the other rows that correspond to the same state).

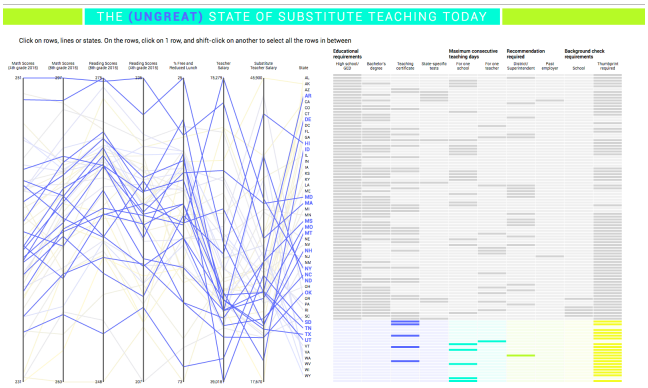
State-specific Information through Information Bar and Highlights on Visualization

The information bar at the bottom appears when users select a particular state or permit row, and disappears upon clicking again. It provides the information displayed in the matrix and parallel coordinates in a condensed and more detailed form (e.g. additional quantitative information like the maximum number of consecutive days is also included). A bottom bar was chosen so that this data is still understood within the whole visualization, and so that matrix column labels are still visible. This feature is particularly relevant for prospective substitute teachers who are interested to quickly locate state-specific information.

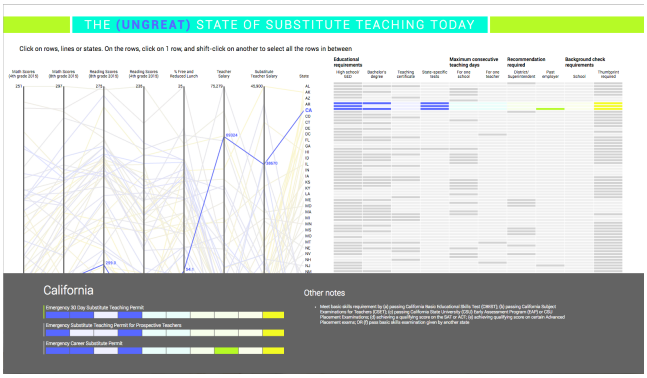
While we initially wanted to include quick links to state substitute-teaching websites, we realized that application to become a substitute teacher was district-specific. Given that our data focuses only on state information, we opted to leave out this actionable step for clarity and consistency.



4.1. Initial display on load



4.2. Sorted matrix with multiple rows selected and corresponding lines highlighted in parallel coordinates



4.3. Single state (California) selected with details in information bar for each permit

Figure 4. Selected components of the final visualization

RESULTS

Description of Informal User Study

Since the task flow for the second user group is more complex than the first, we tested the application with 3 users, all of whom fall into the second category of the general public who might be interested to learn more about substitute teaching. The primary objective of the user study was to determine the ease of determining new insights from exploring the data on this application.

We informally tested the application with 8 participants from a variety of backgrounds (students, working professionals not in the field of education, and members of the data visualization community), during the project fair. We explained to users the purpose of the visualization, encouraged them to explore the application, and asked them for their feedback on confusing portions, and interesting new features.

User Evaluation Results

Participants spent an average of 2 minutes interacting with the visualization, and had generally positive feedback. Participants particularly appreciated the ability to select a specific state, and found that helpful in gleaning useful state-level insights.

In terms of the usability, users picked up 2 major usability issues. First, participants found it difficult to select multiple rows – there was no explicit instruction explaining the "Shift-Click" method of row selection. Second, the reading and math scores column did not have the word "scores" and thus did not make sense. These have since been fixed.

Participants also suggested a few useful features that we have implemented. First, creating a color gradient for parallel coordinate lines, so that a quick glance may yield relationships between multiple features. Second, adding an additional column in the re-orderable matrix to enable sorting based on whether a permit is an emergency permit or not. We implemented the first suggestion, but did not do the second due to time constraints.

In terms of the insights gleaned, many were surprised to learn that there were that many different permits to apply for should one be a substitute teacher. Many participants were disappointed to learn that no clear relationship existed between substitute teaching requirements and other contextual factors, but appreciated that this visualization enabled them to draw this conclusion.

DISCUSSION

Strengths and Limitations of Visualization

This visualization performs very well on the task of gathering and presenting state-specific data within the context of the entire country's state of substitute teaching, and is very suitable for a targeted task flow the first target group might be interested in.

It has also successfully taken two highly multivariate data sets (with a total of 18 variables) and visualized them in an elegant way that can be consumed at a glance, supporting Schneiderman's task taxonomy effectively through incorporating Tufte's principles for effective high-resolution data graphics [17].

However, users have found it challenging to glean any useful and actionable insights from the data exploration process. For instance, when one brushes on the rows that do not require a high school certificate to be a substitute

teacher, the selected parallel coordinates do not show any substantive relationship in any column.

It is possible that we have not chosen the appropriate variables to visualize, and that picking more appropriate contextual information might allow us to glean better relationships. Another possibility is that the contextual data is not granular enough – if data related to substitute teaching salary and test scores existed on a permit-level instead of a state-level, we might have been able to draw stronger conclusions. However, it could very well be the case that the lack of insight is the insight itself. For instance, a possible insight is that the quality of substitute teaching might have little bearing on academic achievement, or that substitute teaching salary has little correlation to how educated the teacher is.

Contribution and Future Work

Prior to this work, there has been no known visual comparison of state-level data on substitute teaching. The visualization serves as the first tool available for prospective substitute teachers and the general public to understand more about substitute teaching on a state-by-state comparison and state-specific level.

One direction of future work is to refine and extend existing features and visualization components, from enabling multi-dimensional brushing of parallel coordinates (a staple of many parallel coordinate visualizations) to defining stronger principles for row clustering in the re-orderable matrix (e.g. weighting particular columns by priority in order). Further work also needs to be done in evaluating the usability of the current application, through a more thorough heuristic evaluation or quantitative user study.

This is also the first time known to the authors where two different but semantically relevant data sets related by only 1 common variable (i.e. state) are displayed together in a parallel-coordinate / re-orderable matrix combination. Previous such visualizations [11] displayed the same data set in both components. More work need to be done to evaluate the effectiveness of such a combination, and its potential to be used in other domains and applications as well.

CONCLUSION

We present a novel interactive visualization around the topic of substitute teaching that allows two different groups of users to achieve different task flows. For prospective substitute teachers, this visualization enables them to access state-specific information for their home state. For the general public (including policymakers), it allows them to explore these otherwise convoluted data sets to determine new insights and relationships between substitute teaching requirements and other contextual information (e.g. salaries, poverty levels, academic performance).

Post-project, we plan to make this publicly available as part of Substantial, a non-profit initiative aiming to improve the recruitment, training and support for substitute teachers nationwide.

ACKNOWLEDGMENTS

We thank Jill Vialet, Stanford d.school fellow and founder of Substantial and Playworks, for providing the contextual understanding of the state of substitute teaching nationwide. We are also grateful to Audra Hollingfield for providing and compiling substitute teaching requirement data set. Many thanks to Maneesh Agrawala for guidance and feedback on the initial design, and for teaching this highly informative course.

REFERENCES

1. Status of Substitute Teachers: A State-By-State Summary. *NEA*, 2016. http://www.nea.org/home/14813.htm.
2. Miller, R and Murnane, R. Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One Urban School District *Willett Educational Evaluation and Policy Analysis* Vol. 30, No. 2 (Jun., 2008), pp. 181-200 Published by: American Educational Research Association
3. Kronholz, J. No Substitute for a Teacher - Education Next. Education Next, 2013. http://educationnext.org/no-substitute-for-a-teacher
4. Eisenberg, R. Why This Boomer Says It's Time to Disrupt Substitute Teaching. *Next Avenue*, 2016. http://www.nextavenue.org/says-time-disrupt-substitute-teaching/
5. Bertin, J. "Graphique et mathématique: généralisation du traitement graphique de l'information." *Annales. Histoire, Sciences Sociales*. Vol. 24. No. 1. EHESS, 1969.
6. W. S. Robinson. A method for chronologically ordering archaeological deposits, volume 16. *Am Antiq,* 1951.
7. Connor Upton, Gavin Doherty, and Myra ORegan. Usability evaluation of an ecological interface for process control health monitoring. Technical Report TCD-CS-2007-32, Department of Computer Science, Trinity College Dublin, 2007.
8. Siirtola H. and Mäkinen E. "Constructing and Reconstructing the Re-orderable Matrix," *Inf Vis Information Visualization* 4.1 (2005): 32-48.
9. Shneiderman, B. "The eyes have it: a task by data type taxonomy for information visualizations," *Visual Languages, 1996. Proceedings, IEEE Symposium on,* Boulder, CO, 1996, pp. 336-343.
10. Siirtola, H. and Räihä, K. "Interacting with parallel coordinates," *Interact. Comput.* (2006) 18 (6): 1278-1309
11. H. Siirtola, "Combining parallel coordinates with the re-orderable matrix," *Coordinated and Multiple Views in Exploratory Visualization, 2003. Proceedings. International Conference on,* 2003, pp. 63-74.

12. Substitute Teachers. *Bls.gov*, 2016. http://www.bls.gov/oes/current/oes253098.htm
13. NAEP Nations Report Card - State-Level Reading Results. *Nces.ed.gov*, 2016. https://nces.ed.gov/nationsreportcard/reading/stateassessment.aspx.
14. Number and percentage of public school students eligible for free or reduced-price lunch, by state: Selected years, 2000-01 through 2010-11. *National Center for Education Statistics*, 2016. https://nces.ed.gov/programs/digest/d12/tables/dt12_046.asp
15. Bostock, M. Les Misérables Co-occurrence. *Bost.ocks.org*, 2016. https://bost.ocks.org/mike/miserables/.
16. Davies, J. Parallel Coordinates. *Bl.ocks.org*, 2016. https://bl.ocks.org/jasondavies/1341281
17. Edward R. Tufte. 1986. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.